Chapter 3
Methodology: Cluster Analysis of Motivation Variables in the TIMSS Data

Abstract This chapter begins with a description of the IEA's Trends in International Mathematics and Science Study (TIMSS) sampling framework. The research study was based on data from three cycles of TIMSS collected at grades four and eight from 12 jurisdictions (Australia, England, Hong Kong, Hungary, Iran, Japan, Norway, Ontario, Quebec, Singapore, Slovenia, and the USA) that participated at both grades in 1995, 2007, and 2015. The motivation variables available in each cycle of administration are outlined, together with how conceptually similar measures for value, enjoyment, and self-confidence in mathematics were constructed; descriptions of demographic and achievement measures used in the analysis are also provided. Two-step cluster analysis was used to create separate profiles of student motivation for each set of data. The characteristics of each motivational cluster were evaluated to ascertain whether differences in cluster membership were related to student background variables (such as sex, time on homework, parental education, and home resources) and mathematics achievement.

Keywords Cluster analysis · Instrumentation · Motivation scales · TIMSS sampling

3.1   TIMSS Sampling

TIMSS is an international study of the mathematics and science performance of students at grades four and eight. Starting in 1995, and conducted every four years since then, TIMSS has collected data from multiple countries; more than 60 countries[1] or jurisdictions, and more than 580,000 students participated in the 2015 cycle of assessment. As well as information on mathematics and science performance, the databases include data from context and background questionnaires completed by the students, their teachers, and their parents.

---

[1]The educational systems examined in our analyses are usually referred to as "countries." This is for ease of reading, but it should be noted that there are a number of systems that are not countries as such, but are units with a degree of educational autonomy that have participated in TIMSS following the same standards for sampling and testing.

In 1995, data were collected from three target populations in 45 countries. These were defined as (a) the two adjacent grades where the majority of nine-year-old students were enrolled, (b) the two adjacent grades where the majority of 13-year-old students were enrolled, and (c) students in their final year of secondary education (Martin and Kelly 1996). In 1999, the target population was limited to grade eight students. From 2003 onwards, the sampling scheme has included students in their fourth and eighth year of schooling, while, in 2015, students in their last year of high school were also sampled.

To select a sample that is representative of the population of students in each participating country, a two-stage random sampling design is used (LaRoche et al. 2016). In the first stage, schools are sampled from each national school sampling frame (a list of all schools with eligible students) with probabilities proportional to school size, and may be stratified by important demographic variables. Once the number of sampled schools is determined in each explicit stratum, systematic sampling proportional to size is used to select schools in each stratum. Provisions for replacement schools are also made. In the second stage, intact classes are chosen through equal probability systematic random sampling. Hence, there is a multilevel structure, where the students are nested in classrooms, and the classrooms nested in schools in each country or jurisdiction. In the 2015 administration, for the sampling precision that is usually required, 150 schools had to be selected for most participating countries, and a sample of 4000 students in each grade (LaRoche et al. 2016).

3.2   Jurisdictions Included in This Study

The countries and benchmarking participants included in the current study were those which participated in the first (1995), the last (2015), and an intermediate administration (2007), provided that they had not been flagged, and that their data are comparable for measuring trends to 2015 (Mullis et al. 2016, Appendix A). The countries which fulfilled these criteria were Australia, England, Hong Kong, Hungary, Iran, Japan, Singapore, Slovenia, and the USA, as well as Norway (grades four and eight), and Ontario and Quebec (Canada), which served as benchmarking participants. These twelve jurisdictions have participated in all TIMSS cycles at both grades four and eight.

Note that, in 1995, data for Ontario and Quebec were obtained as part of a Canadian sample. It was possible to identify schools in those two provinces from the Canada data file using the appropriate school codes (P. Foy, personal communication, 9 August 2018). These two provinces were oversampled in subsequent TIMSS administrations, which makes their results comparable to those of other countries and benchmarking participants.

For the 12 countries and benchmarking participants selected as our sample, the number of grade four students (a fraction of population 1) participating in TIMSS 1995 in each country ranged from 723 to 7296 (see Table 3.1). The range was somewhat smaller for grade eight students (1059 to 7392, a fraction of population 2). Ontario (Canada) had the smallest sample sizes at both grades. All other jurisdictions had much larger sample sizes.

Table 3.1 Sample sizes for the countries and benchmarking participants analyzed (valid percentage of girls)

Participating jurisdictions	TIMSS 1995				TIMSS 2007		TIMSS 2015	
	Population 1[a] students	Grade 4 students	Population 2[a] students	Grade 8 students	Grade 4 students	Grade 8 students	Grade 4 students	Grade 8 students
Countries								
Australia	11,248	6065 (49.9)	12,852	7392 (51.4)	4108 (50.0)	4069 (45.3)	6057 (48.9)	10,338 (50.5)
England[b]	6182	3126 (50.6)	3579	1776 (48.0)	4316 (50.0)	4025 (51.8)	4006 (50.6)	4814 (50.7)
Hong Kong	8807	4411 (45.9)	6752	3339 (45.2)	3791 (48.5)	3470 (50.4)	3600 (44.9)	4155 (47.5)
Hungary	6044	3006 (49.8)	5978	2912 (51.1)	4048 (49.7)	4111 (49.9)	5036 (49.8)	4893 (50.6)
Iran	6746	3385 (48.9)	7429	3694 (44.5)	3833 (47.2)	3981 (44.9)	3823 (48.7)	6130 (48.9)
Japan	8612	4306 (50.0)	10,271	5141 (48.5)	4487 (49.3)	4312 (49.7)	4383 (50.2)	4745 (51.0)
Singapore	14,169	7139 (47.4)	8285	4644 (49.7)	5041 (49.2)	4599 (48.8)	6517 (48.8)	6116 (48.7)
Slovenia	5087	2566 (50.5)	5606	2708 (51.1)	4351 (49.5)	4043 (50.0)	4445 (48.4)	4257 (48.2)
USA	11,115	7296 (51.4)	10,973	7087 (50.2)	7896 (51.0)	7377 (50.4)	10,029 (50.6)	10,221 (50.1)
Benchmarking participants								
Norway	4476	–[c]	5736	–[c]	4108 (49.4)	4627 (49.5)	4164 (49.4)	4795 (50.1)
Ontario	1416	723 (45.6)	2078	1059 (49.7)	3496 (49.3)	3448 (50.6)	4574 (48.2)	4520 (49.8)
Quebec	8470	4488 (50.4)	8378	4245 (50.0)	3885 (51.4)	3956 (49.5)	2798 (50.0)	3950 (52.3)

Notes

[a] In 1995 the target populations included: Population 1, defined as "[s]tudents enrolled in the two adjacent grades that contained the largest proportion of 9-year-old students at the time of testing—third- and fourth-grade students in most countries;" and Population 2, which comprised "[s]tudents enrolled in the two adjacent grades that contain the largest proportion of 13-year-olds at the time of testing—seventh- and eighth-grade students in most countries" (Gonzales and Smith 1997, pp. 1–4)

[b] In England, year 5 and year 9 are considered equivalent to grades 4 and 8, respectively

[c] In Norway in 1995, only grades 2–3, and grades 6–7 participated, so the two samples were excluded from the analysis

The sample of students participating in TIMSS 2007 was roughly similar in most countries (3448 to 5041 students per jurisdiction) with the exception of the USA, which was the only country with a larger sample of participating students (7896 grade four students and 7377 grade eight students). The number of grade four students from the 12 selected countries participating in TIMSS 2015 ranged from 2798 in Quebec (Canada) to 10,029 in the USA. Finally, the number of grade eight students participating in TIMSS 2015 ranged from 3950 in Quebec (Canada) to 10,221 in the USA. The numbers of students in each sample and administration (see Table 3.1) were sufficient (perhaps with the exception of the Ontario province in the 1995 administration) to allow robust generalizations about populations within each jurisdiction.

3.3 Instrumentation

The TIMSS background questionnaires collect information related to attitudes, motivation, and affect in the study of mathematics. However, there is no solid theoretical background underlying the selection of items that are included in the questionnaires. Reviewing the documentation of the questionnaires, there seems to be a gradual tendency towards a more theoretically-justified selection of items and scales across time. In 1995, the theoretical background did not make reference to motivational theories, and the various items were administered as single indicators; a few items could be grouped into an overall "attitude" scale for both grades (plus a "values" scale for grade eight). In contrast, by the 2015 cycle of TIMSS, the theoretical framework made reference to psychological constructs, including specific motivational variables such as enjoyment, value, and confidence in mathematics; each one was operationalized in a separate, multiple-item scale.

In our analysis, we have taken a construct-level approach: beginning with the latest administration, we extract the relevant motivational variables. Then we attempt to trace items that could represent those constructs in the earlier administrations by relying on TIMSS documentation, item content, and on empirical, factor-analytic evidence. We used exploratory factor analysis, with principal axis estimation and oblique rotation, to validate factor structures where these have not been explicitly reported in the TIMSS manuals. The number of factors was determined by reference to factors having eigenvalues >1 (Kaiser criterion) and by examination of the elbow in scree plots. A factor was accepted if item loadings on the expected factor exceeded 0.30 and cross-loadings were less than 0.30 (Bandalos and Finney 2010).

3.3.1 Motivation Measures in the TIMSS 2015 Administration

Hooper et al. (2013) used Deci and Ryan's (1985) theory of motivation to describe the construct of motivation used in the TIMSS 2015 assessment framework. This theory distinguishes intrinsic from extrinsic motivation, as explained in Chap. 2. In TIMSS 2015, students' enjoyment and interest in learning mathematics was measured with nine four-point Likert-type[2] items (see Table 3.2). From these nine items a scale variable termed "Students like learning mathematics" was derived, which we also used for our analyses. Confidence in learning mathematics was measured by nine TIMSS questionnaire items, which were used to derive the scale termed "Student confidence in mathematics" (Table 3.2). For grade eight students, nine four-point Likert-type rating items (Table 3.2) were administered to derive a scale to capture the value component. The scale variable derived from these nine items was termed "Students value mathematics" (Hooper et al. 2013).

In the TIMSS 2015 data, context subscales were scaled using the Rasch partial credit item response theory (IRT) model (Masters 1982); corresponding variables are available as described in Martin et al. (2016b). Using the combined data from all participating countries, each item's model parameters were estimated. Subsequently, individual scores for each respondent were computed, ranging from approximately −5 to 5, and then transformed to a scale that had a mean of 10 and a standard deviation of 2 across all countries. The continuous scales for enjoyment, confidence, and value were used in our analyses.

3.3.2 Motivation Measures in the TIMSS 2007 Administration

In the assessment framework for TIMSS 2007, Mullis et al. (2005) described student motivation as a construct involving students' enjoyment of a subject, values students placed on a particular subject, and their perceived importance of a subject. Student self-concept in mathematics is also considered to influence students' motivation. Student motivation in TIMSS 2007 was measured with seven four-point Likert-type scale items derived from the student background questionnaire (Table 3.3). An additional scale measuring students' value of mathematics, which consisted of four four-point Likert-type scale items, was included in the grade eight student background questionnaire (Foy and Olson 2009).

In TIMSS 2007, items were grouped under three constructs and index variables were calculated for self-confidence (four items), positive affect (three items), and valuing mathematics (four items). However, no scaling was conducted, and unlike the

---

[2]In TIMSS, students are asked to indicate their degree of agreement with a number of statements. The response categories for all motivation items are: Agree a lot, Agree a little, Disagree a little, and Disagree a lot.

Table 3.2 TIMSS 2015 questionnaire items used to measure students' enjoyment, confidence, and value

Enjoyment: Students like learning mathematics questionnaire items
(a) I enjoy learning mathematics
(b) I wish I did not have to study mathematics
(c) Mathematics is boring
(d) I learn many interesting things in mathematics
(e) I like mathematics
(f) I like any schoolwork that involves numbers
(g) I like to solve mathematics problems
(h) I look forward to mathematics lessons ("class" instead of "lesson" for grade 8)
(i) Mathematics is one of my favorite subjects
Confidence: Student confidence in mathematics questionnaire items
(a) I usually do well in mathematics
(b) Mathematics is more difficult for me than for many of my classmates
(c) I am just not good at mathematics ("Mathematics is not one of my strengths" for grade 8)
(d) I learn things quickly in mathematics
(e) Mathematics makes me nervous
(f) I am good at working out difficult mathematics problems
(g) My teacher tells me I am good at mathematics
(h) Mathematics is harder for me than any other subject
(i) Mathematics makes me confused
Value: Students value mathematics questionnaire items (grade 8 only)
(a) I think learning mathematics will help me in my daily life
(b) I need mathematics to learn other school subjects
(c) I need to do well in mathematics to get into the <university> of my choice
(d) I need to do well in mathematics to get the job I want
(e) I would like a job that involves using mathematics
(f) It is important to learn about mathematics to get ahead in the world
(g) Learning mathematics will give me more job opportunities when I am an adult
(h) My parents think that it is important that I do well in mathematics
(i) It is important to do well in mathematics

Rasch-scaled variables of the TIMSS 2015 administration, there were no continuous scales for the motivational variables. Therefore, we investigated whether there was empirical support for grouping and averaging items together to create new variables for confidence, enjoyment, and value for mathematics.

At grade four, TIMSS 2007 included motivation scales in the student background questionnaire that were designed to measure their confidence and affect in mathematics (four and three items, respectively; see Table 3.3). We conducted

Table 3.3 TIMSS 2007 questionnaire items used to measure students' confidence, enjoyment, and value

Confidence: Self-confidence in learning mathematics questionnaire items
(a) I usually do well in mathematics
(b) Mathematics is harder for me than for many of my classmates (at grade 8 this was "more difficult" instead of "harder")
(c) I am just not good at mathematics (at grade 8 this was "mathematics is not one of my strengths")
(d) I learn things quickly in mathematics
Enjoyment: Students positive affect (enjoyment) toward mathematics questionnaire items
(a) I enjoy learning mathematics
(b) Mathematics is boring
(c) I like mathematics
Value: Student valuing mathematics questionnaire items (grade 8 only)
(a) I think learning mathematics will help me in my daily life
(b) I need mathematics to learn other school subjects
(c) I need to do well in mathematics to get into the \<university\> of my choice
(d) I need to do well in mathematics to get the job I want

Notes The item "I would like to do more mathematics in school" was administered but not assigned to an index in the TIMSS 2007 framework, so it was not considered in the analysis

exploratory factor analysis (EFA) on each country's sample using principal axis factoring and oblique rotation. In 10 of the samples, two factors were extracted with the Kaiser criterion and explained more than 61% of the variance. The items loaded strongly on their respective factors, with no cross-loadings above 0.30. In two of the samples (Iran and Japan), one factor was extracted using the Kaiser criterion, although the scree plot was ambivalent. Overall, we interpreted this as evidence that these item groups measured two constructs, and that the two sets of items could be combined to create scores for confidence in and enjoyment of mathematics.

We followed a similar approach for grade eight samples. Here, 11 items were included to capture confidence, enjoyment, and value for mathematics (four, three, and four items, respectively). With a principal axis factoring and oblique rotation, nine of the 12 samples resulted in a three-factor solution, as anticipated. At least 63% of the variance was explained by the extracted factors and no cross-loadings above 0.30 were found. In Hong Kong, Iran, and Singapore, two factors were extracted: one comprised the value items and the other comprised of the enjoyment and confidence items. In the Japanese sample, three factors were extracted; however, two of the

value items ("I think learning mathematics will help me in my daily life" and "I need mathematics to learn other school subjects") cross-loaded on both the value and enjoyment factors.

Since in most of the EFAs item responses loaded onto their intended factors, we created two new variables for grade four and three for grade eight by averaging items (according to the groupings in Table 3.3). If, for an individual student, two or more items responses were missing, we specified the average score as a missing value.

3.3.3 Motivation Measures in the 1995 Administration

No assessment framework existed for TIMSS 1995; however, in the TIMSS 1995 technical report, Martin and Kelly (1996) stated that students' interest, motivation, and effort were merged in a single construct since they were hard to distinguish from each other. Items that measured students' values, competence, enjoyment, interest, and importance (five and 12 items in the grade four and eight questionnaires, respectively) were considered to reflect students' reported motivation for learning mathematics (Table 3.4).

Table 3.4 TIMSS 1995 questionnaire items used to measure students' confidence, enjoyment, and value	*Confidence: Self-confidence in learning mathematics*
	(a) How well do you usually do in math \<s> (at grade 8 this was "I usually do well in mathematics")
	(b) Math \<s> is an easy subject (at grade 8 this was "mathematics")
	Enjoyment: Enjoyment of mathematics
	(a) How much do you like math \<s> (at grade 8 this was "mathematics")
	(b) I enjoy learning math \<s> (at grade 8 this was "mathematics")
	(c) Math \<s> is boring (at grade 8 this was "mathematics")
	Value: Value in learning mathematics (included at grade 8 only)
	(a) I think it is important to do well in mathematics at school
	(b) Mathematics is important to everyone's life
	(c) I would like a job that involved using mathematics
	(d) I need to do well in mathematics… to get the job I want
	(e) I need to do well in mathematics… to get into the \<secondary school> or university I prefer
	Notes Items "I need to do well in mathematics… to please my parents" and "I need to do well in mathematics… to please myself" were not included in our calculation of the value in learning mathematics variable

Since the theoretical framework did not describe specific factors beyond general attitudes for mathematics, we conducted parallel analysis for each country dataset to determine the number of motivational factors. Principal axis factoring with oblique rotation and fixed number of factors followed to examine which of the five ordinal items that were administered to grade four students loaded on each factor. For grade four, results showed that items loaded on two motivational factors for all the countries included in the study, except for Iran. Specifically, three items loaded on one factor (i.e., enjoyment) and two on another one (i.e., confidence), based on the examination of the items' contents. Cross-loadings >0.30 were only observed in Hungary, where two of the enjoyment items loaded on both of the factors. In this case, the primary loading coefficients were taken into consideration when determining the factor structure. A single factor was extracted for Iran, and one of the items had a near zero loading.

For grade eight, the findings were more complex. We conducted parallel analysis for each country to determine the number of motivational factors. Results showed that items loaded on four or five motivational factors for most of the countries included in the project (except for Iran). Principal axis factoring with oblique rotation and fixed number of factors followed, to examine which items loaded on each factor. For most countries (except for Iran and Hungary), three items loaded on one factor measuring enjoyment, and two on a confidence factor. In Iran, all five items loaded on a single factor, while in Hungary the two items measuring students' confidence in mathematics loaded onto two different factors. The remaining seven items loaded on two or three factors. The items "I need to do well in mathematics to please my parents" and "I need to do well in mathematics to please myself" formed a factor or were the single items loading on a factor (for seven out of the 11 countries included in our analysis). Because those two items were not included in the TIMSS 2015 administration, we excluded them from further analyses. The remaining five items usually loaded inconsistently onto two factors. Since these items were included in one scale named "value" in TIMSS 2015, we also considered these as forming a single scale in TIMSS 1995 (see Table 3.4). After assessing the results of our EFA, we created variables by averaging items. If for an individual student, two or more item responses were missing, we specified the average score as a missing value.

It is worth noting that, due to local considerations, various nations may not have administered certain items in certain rounds.[3] For example, the item "I think it is important to do well in mathematics at school" (Table 3.4) was not administered in Norway in 1995. Nonetheless, the scale score for the construct was calculated for those nations using the remaining items.

---

[3]For more detailed information about individual national adaptations undertaken in the TIMSS 2007 and 2015 administrations, see Foy and Olson (2009) and Foy (2017, tables 2 and 3, and supplement 2).

3.4 Other Variables Included in the Study

3.4.1 TIMSS Achievement Score Estimation

To ensure adequate content coverage, a large pool of assessment items is administered in each cycle of TIMSS. The burden of responding to hundreds of questions would be too great for any student, so TIMSS uses a planned missing data, multiple-matrix sampling. Each examinee receives a subset of the item pool. In this way, individual student testing time is reasonable, while it is possible to obtain measures of performance on broad content domains at the aggregate level (Rutkowski et al. 2014). When estimating the distribution of proficiencies for the large population of students, there are problems of inaccuracy in the resulting estimates of individual proficiencies and population characteristics. Plausible values (PVs) are employed to address this problem: all the data, including student responses and their background data are used to estimate PVs. PVs are multiple imputed scores from the estimated conditional ability distributions (given all the students' responses and background data). They can be thought of as imputed scores for "students with similar response patterns and background characteristics in the sampled population" (Martin et al. 2016a, p. 12.5), and aim to present estimates of population parameters; they are not used as estimates of individual student scores. The five PVs estimated for each student are representative of the variation in estimating individual student proficiency (Martin et al. 2016a).

Achievement data were scaled through IRT models, which are latent variable models that estimate the probability of a specific student answering items correctly based on the student's proficiency, which is the latent trait θ. For dichotomous response questions (e.g., multiple-choice items marked as correct or incorrect) the three-parameter IRT model was used; this model accounts for three parameters: item difficulty, item discrimination, and item pseudo-chance. For constructed response items, which do not have options from which to select but which are also marked as correct or incorrect, a two-parameter IRT model with parameters for item discrimination and difficulty was employed. For polytomous items, the partial credit model was used (Martin et al. 2016a).

International achievement in TIMSS 1995 was reported using only one plausible value, however all five PVs are available in the TIMSS 1995 data (Gonzales and Smith 1997). TIMSS cycles after 1995 reported five PVs and we used these as indicators of students' achievement in mathematics. When we report the average performance for a student group, all five PVs were considered and total student weights were applied using the IEA's International Database Analyzer (IDB) software (see www.iea.nl/data for further information about this free-to-download analysis tool).

3.4.2 Other Variables of Interest

Student Sex The students' sex variable was used in the present study.

Time on Homework Items measuring the number of hours spent on studying or doing homework differ across TIMSS administrations. TIMSS 1995 student questionnaire included the question "On a normal school day, how much time before or after school do you spend doing each of these things? […] studying math or doing math homework after school," which measured the time a student spent on studying mathematics. Two questions examining the time spent on mathematics homework ("How often does your teacher give you homework in mathematics?" and "When your teacher gives you mathematics homework, about how many minutes do you usually spend on your homework?") were included in the TIMSS 2007 and TIMSS 2015 student questionnaires. An index variable, consisting of three categories, combined the two responses in the 2007 dataset. A derived variable with three categories was extracted from the two items; this measured the weekly time a student spent on mathematics homework. The time spent on mathematics homework item was not included in the TIMSS 2015 grade four student questionnaire. All items measuring time spent on studying/doing homework in the three TIMSS administrations that we considered (1995, 2007, and 2015) contained five response options.

Parental Education Across all TIMSS administrations, the grade eight student questionnaire includes questions asking students to report the highest level of parental education. Parental education questions were omitted from the grade four student questionnaires. The items measuring parental educational level are highly similar across TIMSS administrations, however the response options vary across administrations and across nations. At grade eight, in all three of the TIMSS administrations that we selected for our study, two separate items asked for the highest education level achieved by mother and father. There were eight response options in 2015 and 2007, and just seven response options in 1995. A derived variable with six categories of education levels in 2015 and 2007, and four categories of education level in 1995, was created by combining the observed variables about the father and mother into a parental education variable. Here we report the percentage of parents with education above a cut point: "above secondary" for TIMSS 1995 and "post-secondary and above" for TIMSS 2007. We did not analyze this variable for TIMSS 2015 because it was included in the definition of the more comprehensive home resources variables.

Home resources Home resources are proxy variables for a student's socioeconomic status (SES), and are only available in the TIMSS 2015 data. Items for parental education level, occupation, income, and number of books in home are used as indicators of students' SES. There are two derived scale variables in TIMSS 2015: the "Home resources for learning" scale for grade four students and the "Home educational resources" scale for grade eight students. The two derived scales include number of books and number of children's books at home, number of home study supports (own room and internet connection) and parental educational and

occupational level (grade four) or number of books in the home, number of home study supports (own room and internet connection), and parental educational level (grade eight) (Martin et al. 2016b).

These two scales were calculated for TIMSS 2015, but not for TIMSS 2007 and 1995, because relevant items have not been surveyed consistently. Hence, we were unable to include proxies for SES in the analyses of earlier administrations.

3.5 Analysis Technique

In cluster analysis, similar observations in a dataset are grouped together in a cluster (Bartholomew et al. 2008). Similarity is determined by information from one or more of the variable characteristics of the observations. The grouping is not known in advance. Identification of homogenous observations is essentially a taxonomy analysis.

While there are several techniques for cluster analysis, we here outline three common approaches that are available in statistical packages (such as the IBM SPSS Statistics package; see https://www.ibm.com/support/knowledgecenter/en/SSLVMB_24.0.0/spss/base/cluster_choosing.html). Firstly, hierarchical cluster analysis is an agglomerative procedure that begins with each observation as a separate group, and gradually combines observations or groups based on similarity, until one large cluster is formed. The hierarchical approach is recommended when input variables are continuous and the sample of observations is small. A dendrogram is produced and examined to ascertain the number of clusters to retain and their meaning. K-means clustering can be used with continuous variables and large datasets. The number of clusters must be defined in advanced. Solutions with different numbers of clusters can be inspected and compared. Finally, two-step cluster analysis can handle both continuous and categorical variables in very large datasets to generate a solution; first by constructing a cluster features tree to summarize the observations and then by employing an agglomerative algorithm.

Because cluster analysis is an exploratory procedure, different numbers of clusters may be extracted and interpreted, especially when using two-step or K-means clustering. In our preliminary analyses, a small number of clusters were extracted (e.g., two or three). In these solutions, the clusters were consistent and not very informative with respect to the input variables. For example, one cluster was composed of students with high scores on all input variables, another cluster grouped the students with moderate scores, and a third cluster was composed of students with rather low motivation scores. This approach did not permit the identification of possible inconsistent profiles across the motivational constructs, which was an important aim of our study.

Therefore, within a two-step cluster approach, the fixed number of clusters was incremented between three and five at grade four and between three and six at grade eight; more clusters were examined for grade eight because one additional input variable ("Value for mathematics") was available for these older students. These

numbers were selected so that (a) the analysis would produce more than just clusters with consistent motivation responses, (b) a manageable number of reasonably-sized clusters would be produced, increasing the likelihood they would cross-validate, and (c) inconsistent motivational profiles could be identified. Clusters with students scoring medium-to-low on one input variable and high on another input variable present potential theoretically interesting opportunities. For instance, students who value, but do not necessarily like mathematics, or students with high self-confidence in the subject, but who report low value and enjoyment in mathematics classes, may have more or less successful achievement profiles or differ on sociodemographic characteristics. Cluster comparisons on evaluation variables such as achievement and demographics offer insights into the possible predictors or outcomes of such inconsistent motivational profiles.

Due to the exploratory nature of cluster analysis, the evaluation of competing cluster solutions was not automatically determined. In choosing the final solution, we considered statistical measures, such as the silhouette measure of cohesion and separation (at least "fair"; Kaufman and Rousseeuw 1990), and the relative size of the smallest cluster (>7% of the sample). In addition, we considered the interpretation of the derived clusters. The final number of clusters for each country sample, in each cycle of TIMSS (2015, 2007, and 1995), and at each grade (four and eight) was decided based on the assessment of two independent researchers. When agreement could not be reached, a decision was adjudicated in the presence of a third researcher.

The two-step cluster method was implemented for each sample. This approach is available in the IBM SPSS Statistics software and is appropriate for large samples (Appendix A provides the syntax used to generate clusters). The following measures were entered as input variables:

(1) Students like learning mathematics/enjoyment
(2) Students confident in mathematics/confidence
(3) Students value mathematics/value (this applies only for grade eight).

These scale variables were available in the TIMSS 2015 administration datasets and were derived from context questionnaire items using IRT procedures. For the TIMSS 2007 and 1995 administrations, the variables used were derived after theoretical considerations and factor analytic procedures, by averaging the relevant items as described in the instrumentation section. Hence, the motivation scores in TIMSS 2015 are expressed on a different scale from that used for the earlier administrations.

Because the results of the clustering algorithm in SPSS depend on the order of cases in a dataset, prior to each analysis we undertook the following steps: (a) the cases of each dataset were sorted by each student's unique ID, (b) a fixed seed was specified, and (c) an observation was generated from a continuous uniform distribution for each case for a random ordering of the dataset. Thus, each cluster solution was based on the same reproducible ordering of cases. The following code was always added before any clustering procedure:

```
sort cases by IDSTUD(A).
set rng mc seed 123456789.
compute randvar=rv.uniform(1,1000).
```

sort cases by randvar.
delete variables randvar.

Because the TIMSS data has a nested structure, we here note that the literature identifies two approaches for dealing with sampling weights. The design-based approach recommends using the sampling weights in order to avoid biased parameter estimates. Conversely, the model-based approach does not suggest the use of sampling weights because, if the correct ("true") model is specified, the use of weights leads to a decrease in efficiency and precision (Anderson et al. 2014; Snijders and Bosker 2012). The IBM SPSS Statistics two-step cluster procedure also does not permit the use of sampling weights and ignores the specification on the WEIGHT command.[4] Thus, we did not use sampling weights in our cluster analyses. With respect to missing values, cases were excluded from the cluster analysis when a value was missing from the input variables.

When evaluating the clusters, we examined the following background and achievement variables for each cluster:

(1) Average performance in mathematics (PVs 1 through 5)
(2) Percentage of girls in the cluster
(3) Percentage of students with a high level of parental education (applies only for grade eight students, and administrations 2007 and 1995)
(4) Home resources: the average "home resources for learning" (only available for TIMSS 2015 grade four students) and "home educational resources" (only available for TIMSS 2015 grade eight students), as indications of SES
(5) Time spent on homework, with the caveats:

  • TIMSS 2015: This was not available for grade four students; at grade eight, TIMSS 2015 reported the percentage of students that spent "> 45 min on homework weekly."
  • TIMSS 2007: The variable used was the "index of time on math homework."
  • TIMSS 1995: TIMSS 1995 reported the percentage of students that spent "more than 1 h on homework daily."

Finally, we conducted statistical tests for comparing cluster means. We undertook pairwise mean comparisons using the IEA's IDB Analyzer, which allowed us to estimate weighted statistics and corrected standard errors for all the TIMSS assessments. Clusters were compared on average performance in mathematics using all five plausible values for all administrations and samples, and, for 2015, on the home resources variables. Since multiple pairwise tests were conducted for each jurisdiction, we adopted an alpha level of 0.001; a difference was considered statistically significant if the t-statistic (in absolute value) exceeded 3.29. We employed chi-square testing to examine dependencies between gender and clusters. Parental education and homework engagement were measured with different response scales across samples; we report descriptive statistics by cluster for those two variables.

---

[4]https://www.ibm.com/support/knowledgecenter/en/SSLVMB_24.0.0/spss/base/syn_twostep_cluster_overview.html.

References

Anderson, C. J., Kim, J. S., & Keller, B. (2014). Multilevel modeling of categorical response variables. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 481–519). Boca Raton, FL: CRC Press.

Bandalos, D. L., & Finney, S. J. (2010). Factor analysis: Exploratory and confirmatory. In G. R. Hancock, & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 93–114). New York, NY: Routledge.

Bartholomew, D. J., Steele, F., Galbraith, J., & Moustaki, I. (2008). *Analysis of multivariate social science data*. Boca Raton, FL: Chapman and Hall/CRC.

Deci, E., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York, NY: Plenum Press.

Foy, P. (2017). *TIMSS 2015 user guide for the international database.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timssandpirls.bc.edu/timss2015/international-database/downloads/T15_UserGuide.pdf.

Foy, P., & Olson, J. F. (2009). *TIMSS 2007 user guide for the international database.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Gonzales, E. J., & Smith, T. A. (1997). *User guide for the TIMSS international database. Primary and Middle School Years.* Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Hooper, M., Mullis, I. V. S., & Martin, M. O. (2013). TIMSS 2015 context questionnaire framework. In I. V. S. Mullis, & M. O. Martin (Eds.), *TIMSS 2015 assessment frameworks* (pp. 61–82). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timssandpirls.bc.edu/timss2015/frameworks.html.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York, NY: John Wiley and Sons.

LaRoche, S., Joncas, M., & Foy, P. (2016). Sample design in TIMSS 2015. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015* (pp. 3.1–3.37). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timssandpirls.bc.edu/publications/timss/2015-methods.html.

Martin, M. O., & Kelly, D. L. (Eds.). (1996). *Third International Mathematics and Science Study (TIMSS) technical report, Volume I: Design and development*. Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College. Retrieved from https://timss.bc.edu/timss1995i/TechVol1.html.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (2016a). TIMSS 2015 achievement scaling methodology. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015* (pp. 12.1–12.9). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timssandpirls.bc.edu/publications/timss/2015-methods.html.

Martin, M. O., Mullis, I. V. S., Hooper, M., Yin, L., Foy, P., & Palazzo, L. (2016b). Creating and interpreting the TIMSS 2015 context questionnaire scales. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timssandpirls.bc.edu/publications/timss/2015-methods.html.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timssandpirls.bc.edu/timss2015/international-results/.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2005). *TIMSS 2007 assessment frameworks.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from https://timss.bc.edu/TIMSS2007/frameworks.html.

Rutkowski, L., Gonzales, E., von Davier, M., & Zhou, Y. (2014). Assessment design for international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 75–95). Boca Raton, FL: CRC Press.
Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis* (Second ed.). London, UK: Sage.